What If I Speak Now? A Decision-Theoretic Approach to Personality-Based Turn-Taking

Kathrin Janowski
University of Augsburg
Augsburg, Germany
kathrin.janowski@informatik.uni-augsburg.de

Elisabeth André
University of Augsburg
Augsburg, Germany
andre@informatik.uni-augsburg.de

ABSTRACT

Embodied conversational agents, which are increasingly prevalent in our society, require turn-taking mechanisms that not only generate fluent conversations but are also consistent with the personality and interpersonal stance required in the given context. We present a decision-theoretic approach for deriving the turn-taking behavior of such an agent from the personality it is meant to convey. For this we gathered relevant theories from psychology and communications research, as well as related systems employing utility-based reasoning. On this basis we describe the construction of an influence diagram which decides between acting and waiting based on those actions' expected utility for the agent's personality-related interaction goals. To test our approach, we integrated our model into an application which simulates conversations between two virtual characters. We then evaluated our prototype by presenting videos of those conversations in an online survey. Our results confirmed that differences in an agent's speaking behavior, generated from different Extraversion configurations in our model, lead to the intended perceptions of its Extraversion, Agreeableness and Status.

KEYWORDS

embodied conversational agents; personality modeling; interpersonal stance; turn-taking conflicts; interruptions; decision-theoretic approach; influence diagram; Bayesian network

ACM Reference Format:
Kathrin Janowski and Elisabeth André. 2019. What If I Speak Now? A Decision-Theoretic Approach to Personality-Based Turn-Taking. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Our world is increasingly populated by talking computer-controlled characters. Disembodied voice assistants, such as Apple's Siri or Amazon's Alexa, offer to manage our day-to-day routine. Virtual actors entertain us in video games or coach us for job interviews [10] and cultural sensitivity. And finally, numerous companies promise to bring social robots like Jibo, Buddy or Olly into our homes, advertised as potential family members brimming with personality.

To interact with humans in a natural, intuitive manner, these agents need to know when they are allowed or expected to speak, when to yield their turn, and when to stand their ground in order to deliver a crucial message. Humans use complex mechanisms to negotiate and communicate turn-taking intentions. However, the automatic analysis and generation of those are still active research areas, which makes it difficult to determine the appropriate timing for the system's reaction. Commercially available systems therefore tend to follow a strictly sequential pattern and respond only after the user has been silent for a certain time, which can slow the interaction down. In contrast, systems with incremental speech recognition can react as soon as they can predict the user's intention, enabling a more natural interleaving of request and response [8].

But should the agent really respond as soon as possible? On one hand, interrupting the speaker can be seen as aggressive and undesirable[8]. On the other hand, finishing each other's sentences can also signal understanding [8], interest and involvement in the conversation[12]. Since humans tend to ascribe social characteristics to computer-controlled agents [20], these agents must convey a context-appropriate interpersonal stance. For instance, an agent playing a potential employer in a challenging job interview [10] would be expected to be more dominant and less polite than a personal secretary agent who obediently manages a senior's calendar [14].

This paper provides a decision-theoretic approach for calculating conversational timing based on the agent's personality and the resulting stance towards the interlocutor. Section 2 will outline the psychological background of conversational floor management behaviors, after which section 3 will present existing approaches for modeling them in dialog systems. Section 4 will explain our approach, section 5 will describe its implementation in our first prototype and section 6 will describe its evaluation. Section 7 will summarize this paper and outline future work.

2 PSYCHOLOGICAL BACKGROUND

2.1 Interpersonal Stance and Personality

A person's behavior and attitudes towards others, the so-called interpersonal stance, can be described using the interpersonal circumplex [9, 13, 17]. Its two main axes are usually labeled as *Status*, which ranges from submissive to dominant, and *Love*, which ranges from cold to warm. The latter dimension is also known as *Affiliation* [16], ranging from either hostile or indifferent [13] to friendly.

Alternatively, the interpersonal stance can be expressed via the personality. A wide-spread model for describing personality are the so-called "Big Five" factors [16] which are commonly associated with the following characteristics:

- **Extraversion:** assertive, enthusiastic, energetic, outgoing, talkative and nonverbally expressive.
- **Agreeableness:** forgiving, generous, sympathetic, compassionate and trusting.

© 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). . This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

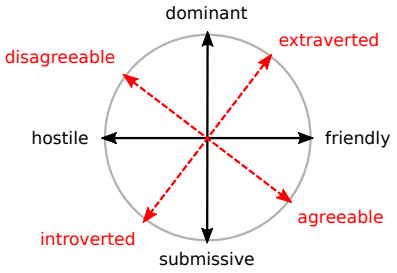

Figure 1: The two pairs of axes used to describe the interpersonal stance. *Solid:* Status and Affiliation. *Dashed:* Extraversion and Agreeableness.

- **Conscientiousness:** thorough, well-organized, responsible, strong-willed, ambitious, productive and adhering to rules.
- **Neuroticism:** likely to experience negative affect, giving in to impulses, having low self-esteem and a hard time coping with problems.
- **Openness:** unconventional thinking, need for variety, curiosity, imagination and a wide range of interests.

McCrae and John [17] and later DeYoung et al. [9] both confirmed that the traits Agreeableness and Extraversion are similarly suited for describing the interpersonal circumplex. Mathematically, the two pairs of axes are rotated by approximately 30° to 45° relative to each other (see figure 1). Dominance thus relates to a combination of high Extraversion and low Agreeableness, while submissive behavior appears introverted and agreeable.

Argyle and Little [2], however, pointed out that a person's behavior and therefore their apparent personality vary between situations, depending on factors such as their current role and present observers. For example, an agreeable person may behave in a friendly and submissive manner towards most people, but be distant and less compliant towards a disliked interlocutor. It is therefore advisable to distinguish between the underlying personality and the interpersonal stance which arises from a combination of personal and situational properties.

2.2 Interpersonal Goals and Speech Overlaps

There are many reasons why speaker intentions may be in conflict. For instance, Rogers and Jones [22] found that highly dominant subjects held the floor for a greater amount of time and made more interruption attempts per minute of the other party's speaking time. However, Goldberg [12] draws attention to other explanations for overlapping speech. Overlaps which are semantically connected to the ongoing sentence can indicate enthusiasm and involvement, showing mutual interest in the topic or shared activity. In this case, a possible motive behind the overlap would be to express affiliation, which (as shown above) is closely related to agreeableness. Other interruptions may arise from a combination of conflicting goals. This can be illustrated with a clarification request, for which the listener may have several intentions:

(1) To understand and give a meaningful response to the speaker, in order to appear competent[12] and therefore of high status[13].
(2) To respect the speaker's goal of receiving an appropriate response[12], which indicates shared communicative goals and therefore affiliation.
(3) To respect the speaker's right of speaking[12], thereby demonstrating submissiveness which conveys low extraversion and/or high agreeableness (see previous section).

Therefore, when the listener wants to start speaking, they need to balance the costs and benefits for these goals against each other [12, 13]. While interrupting the speaker with a question may be detrimental to goal 3, it may be beneficial and even necessary for goals 1 and 2. If the listener waits patiently, but does not understand, they can not respond in the desired way, which would not only irritate the speaker but also harm the listener's own goals.

3 RELATED WORK

3.1 Probabilistic Dialogue Reasoning

One major challenge in dialog systems is dealing with uncertainty, which can arise from sensor noise, semantic ambiguity or non-deterministic user behavior. A common solution is to rely on probabilistic relationships between the system's observations and the true situation. Decision-theoretic approaches then take the possible consequences of every action into account by calculating the utility of their outcomes. This utility depends on the *world state* which holds all the variables defining the present situation, such as each participant's speech activity, their cognitive load or the semantic content that is being spoken. The best actions are then determined by multiplying each world state's utility with its probability.

Bohus and Horvitz [3] modeled the probabilities that the floor was passed to the system and that one of the humans would decide to speak after a certain time, as well as likely delays in the system's input and output processing. They focused on preventing turn-initial overlaps after phases of silence, when both system and user might start speaking at the same time. Based on manually assigned costs for different turn-taking errors, the authors calculated the utility of possible waiting times. When the agent waits longer to take its assigned turn, the expected cost increases due to the risk that a user takes the floor in the meantime. However, it decreases when the roles are reversed, allowing the agent to fill the silence.

Conati [6] described the use of a dynamic Bayesian decision network to infer the user's current emotions and possible causes from knowledge of their goals. Those can depend on various user traits such as their personality. The system then determines the timing for offering a given service that is most beneficial for the user's affective state. The causal relationships modeled in the Bayesian network are used to cope with the uncertainty in the system's perception, in order to decide whether to offer the service now, wait for a better time, or ask the user for help with this decision.

Both works explain the benefits of a decision-theoretic approach for interpersonal coordination, but neither models an explicit personality for the agent or interactional goals tied to their stance towards the human. Bohus and Horvitz described a specific application for a virtual quiz master, while Conati focused on modeling the human's personality and goals. However, Conati points out the usefulness of Bayesian networks for both predictive and diagnostic reasoning, from which we deduce that a similar approach can be used to model the agent's own goals. As for timing conflicts, Bohus

and Horvitz focus on preventing a specific form of overlaps. We intend to generalize their approach to cover others as well. Conati's virtual butler operates on a coarser time scale, for example delaying bad news for a few hours, but we believe that similar mechanisms can also be used for individual phrases.

3.2 Semantically Plausible Interruptions

Interruptions which are semantically related to the current utterance, such as completing each other's sentences or interrupting with a disagreement, are only possible when the interrupter could have plausibly understood the other party's intention.

DeVault, Traum et al. [8, 27] used trained classifiers to incrementally predict the content of the user's finished sentence, and whether this prediction can be improved by listening further. When the latter is unlikely and the user falls silent before finishing, the virtual agent steps in to complete the sentence. To avoid offensive interruptions, this is only allowed after a pause of at least 600 ms. During shorter pauses, the agent provides less intrusive feedback using short backchannel comments or nonverbal signals [27].

Chao [5] calculates the timing in relation to the so-called "minimum necessary information", or "MNI" for short. This term stands for the part of a communicative action which needs to be observed before its meaning becomes evident, such as a keyword or a characteristic gesture stroke. According to Chao, the end of the MNI is a reliable predictor for when a human listener will start responding.

Both works describe the same fundamental idea, both from the agent's perspective and that of the user. We think this MNI principle is very useful for triggering the next dialogue contribution and thus limiting the time frame for a potential interruption. However, instead of using a fixed time threshold as in [27], we intend to adapt the timing based on contextual factors.

3.3 Barge-in Handling

One important part of turn management is the reaction to barge-in interruptions which occur during the speaker's utterance.

The companion agent described by Crook, Smith et al.[7, 25] uses acoustic features and speech duration to distinguish between backchannel comments and full interruptions. In the latter case, it immediately stops speaking and then chooses how to react based on the semantic content of the interrupting phrase. If it detects new information, such as a correction, it re-plans its contribution. Otherwise, it acknowledges the interruption with a short empathetic response and then repeats the interrupted phrase.

Selfridge et al. [23] argue that the system should only yield its turn when the user's speech act is deemed more important. The system's speech is paused while valid user speech is being detected, and resumed if no more is detected during an adaptive time window. When the final recognition result becomes available, the system only reacts to the user's speech if it advances the dialogue. Otherwise, the input is ignored, which means that this approach is not only robust to false-positive speech detection but can also handle backchannel comments in a natural manner.

Both works handle interruptions based on their semantic content, but neither mentions modeling different personalities through the continuation policy. However, both reveal relevant patterns. The empathetic companion yields its turn as soon as possible, conveying an agreeable and submissive personality. The latter system strives to increase task efficiency by allowing the user to cut the system's output short, but also by letting the system keep or quickly re-take the floor when the interruption is deemed irrelevant. This appears more dominant, and might also indicate a conscientious personality. We therefore think that barge-in handling can plausibly be derived from a model of personality and interpersonal stance.

3.4 Personality and Interpersonal Stance

As explained in 2.2, overlapping speech is closely tied to a speaker's personality and stance towards the other participants. This has been confirmed in perception studies with virtual agents.

Ter Maat et al. [26] defined several turn-taking and -yielding heuristics for conversations between two agents. They found that starting to speak before the end of the other's turn was perceived as less agreeable than starting afterwards. Agents who waited for a few seconds before responding to a finished turn were rated as less extraverted than those speaking immediately after or before that turn's end. Likewise, yielding the turn in case of overlaps was rated as warmer, less active and less dominant whereas continuing in a louder voice appeared less agreeable and more neurotic.

A similar study by Cafaro et al. [4] further distinguished between disruptive and cooperative overlaps (compare section 2.2). While they do not describe any computational model for generating these behavior patterns, their results showed that the interrupted agent was rated as more dominant and less friendly when it continued speaking for a longer time. However, yielding quickly was perceived as less friendly in case of cooperative utterances, confirming that certain overlaps are desirable. Interrupting while the speaker pauses was perceived as less dominant than causing overlaps, and also more friendly in case of disruptive utterances.

Ravenet et al. [19] modeled their agents' interruption behavior based on the interpersonal stance towards the current speakers. Since dominant persons are more likely to interrupt others and both hostility and friendliness can cause simultaneous speech (see section 2.2), an agent acts on their desire to speak when the sum of their status and absolute affiliation towards the other speaker(s) is greater than zero. This rule controls both the agent's interrupting behavior as well as their reaction to being interrupted themselves.

These works confirm several factors to be considered in a turn-taking model. The perceived interpersonal dynamics not only depend on the duration of the overlap, but also on its semantic content and action parameters such as voice volume. To generate believable behavior, a turn-taking model must therefore be able to reason about numerous interdependent variables while still remaining transparent to the interaction designer. As we will explain, a decision-theoretic approach such as an influence diagram is well-suited for this purpose. Ravenet et al. [19] calculate a time threshold rather than predict potential costs and benefits based on uncertain beliefs. However, we see that their formula can easily be used for the latter, for example based on the probability that a particular person will (continue to) speak. They also provide a means to reduce complexity by using the same rule for both the interrupter's and the interruptee's behavior. Therefore we intend to use a similar simplification in our model.

4 OUR APPROACH

4.1 Choice of Computational Model

An influence diagram is a Bayesian network with added nodes for decisions and the utilities of specific world states [6, 18]. This model has several benefits for modeling turn-taking behavior.

First of all, the Bayesian network can describe observations with more than one explanation, as well as separate observations sharing a common cause. This is ideal for reconciling different interaction goals and cognitive states which, on the surface level, may lead to the same behavior. Likewise, it can describe behavior variations which convey the same meaning.

Second, the decision and utility nodes allow for the calculation of the best policy given uncertain observations. The utility nodes can be used to specify the interaction goals and the events that are beneficial or harmful for them. The magnitude of the utilities can also be scaled to express each goal's priority in relation to the others. The decisions which maximize the total expected utility then represent the optimal behavior choices for attaining that goal.

Furthermore, Bayes' theorem not only lets us infer the most likely explanation for a given observation, but also enables us to predict the observations when their causes are known to be present [6, 18]. This can save a lot of time because a model developed for one side of the conversation can be re-used for the other.

Finally, the graphical structure of an influence diagram can represent causal relationships derived from expert knowledge. This makes the model more transparent and accessible to humans than, for example, artificial neural networks. Nevertheless, machine learning can be used to obtain the probability distributions from annotated data, so this approach offers an elegant way for combining hand-crafted and automatically trained rules.

4.2 Diagram Structure

The perception studies mentioned above confirm the relationship between speech timing and an agent's perceived personality and interpersonal attitude. We therefore take the personality as the starting point for our model. As section 2.1 showed, the interpersonal dimensions *Status* and *Affiliation* are mathematically related to the Big Five factors *Extraversion* and *Agreeableness*, from which we conclude that these traits determine the agent's default stance towards other people. All four are represented as separate chance nodes in our Bayesian network, with conditional dependencies according to the theories in section 2.1. This allows the interaction designer to both configure the agent's personality from which the default interpersonal stance is derived, and to refine the latter based on context information, such as the agent's task-related authority.

The participants' current roles (*speaker* or *listener*) are also modeled as chance nodes. While the agent's role is known, that of the user can only be inferred from observed behaviors which are conditionally dependent on it. The interaction goals are represented by utility nodes, and the actions which affect them are found in the decision nodes. At the very least, the influence diagram needs to decide between *wait* and *act*. More decision nodes could be included for action parameters such as the speech volume, or additional modalities such as the gaze direction.

The first version of our influence diagram, which will be extended with each prototype, can be seen in figure 4. It contains one goal *Exert Control* and one decision for the agent's *own Speech Behavior*, which has the two options *speak* and *wait*. The utility node contains the costs and benefits for both options in all possible combinations of the agent's Status and the situation variables "own Role", "other Role" and "other Speech State Duration". Behavior which is beneficial for the goal, such as speaking over the other's turn, has a positive utility for dominant characters. The magnitude of the utility depends on the relative timing of both agents' speech. For introverted characters, the goal itself is undesirable, which is represented by low or negative utilities for similar behavior.

The total expected utility EU(D) of a decision is calculated as

$$EU(D) = \max_{1 \le j \le m} \sum_{i=1}^{n} P(w_i) * U(d_j, w_i) \tag{1}$$

where $P(w_i)$ is the probability that the world state variables will have a particular combination of values out of the n possible ones when the decision is made, and $U(d_j, w_i)$ is the utility of doing action d_j out of m options given this world state w_i. The action d_j which maximizes the expected utility will then be chosen.

In Figure 4, the personality is set to *(extraverted, neutral)*, causing a high probability for a dominant attitude (81.25%). The agent is still in the *listener* role, and observed that the other party has been speaking for a *short* time, which implies a *speaker's* full turn rather than giving *listener's* feedback. The utility $U(speak)$ regarding goal *Exert Control* depends on *own Status*, *own Role*, *other Role* and *other Speech State Duration*, which make up the world state w_i in this case. The outcomes of *own Role* and *other Speech State Duration* are known and each of the remaining states has two outcomes with $P(outcome) > 0$, which gives us four potential world states to consider:

$$\begin{aligned} EU(speak) &= \sum_{i=1}^{4} P(w_i) * U(speak, w_i) \\ &= (0.1875 * 0.1) * (1.0) + (0.1875 * 0.9) * (-0.4) \\ &+ (0.8125 * 0.1) * (3.0) + (0.8125 * 0.9) * (0.2) \\ &= 0.34125 \end{aligned} \tag{2}$$

5 PROTOTYPE IMPLEMENTATION

Figure 2 shows the relevant components of our dialog system. Our current prototype simulates conversations between two virtual agents[1], following the example of the works in section 3.4. This allows us to reproduce the required interruption episodes without the variability inherent in human behavior, and lets us test the core behavior model without sensor noise.

5.1 Dialog Manager

The script advances based on the MNI (see section 3.2). Our current implementation uses bookmarks in the text-to-speech commands to detect the end of the MNI, at which point it is written to the *Shared Information Board*.

The state machine which models the dialog flow was implemented in Visual SceneMaker [11]. It provides the next utterance as soon as the required MNI was spoken by either participant. At this

[1]Examples of the generated behavior can be seen here: https://www.youtube.com/playlist?list=PLAJ5ZtqkzFRtaO_kK9qPKvjxjzMWawBql

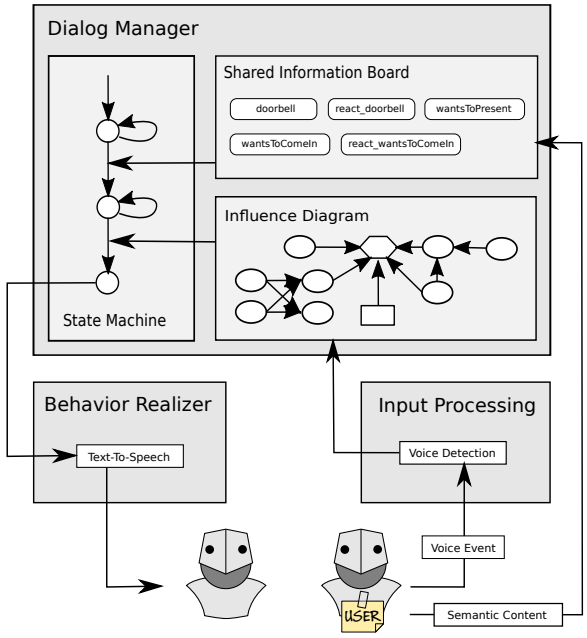

Figure 2: Overview of the current system architecture.

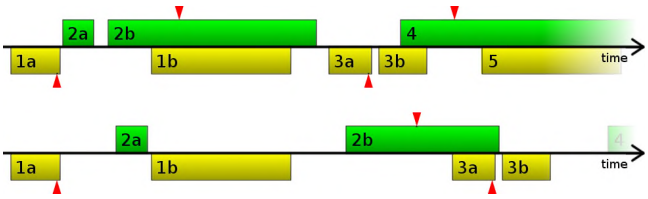

Figure 3: Example timelines for talking to the (extraverted) simulated user. Green bars show the agent's speech, yellow that of the user. The end of the MNI is marked by the triangles. *Top:* Extraverted agent. *Bottom:* Introverted agent.

point, it starts monitoring the influence diagram until the expected utility for speaking exceeds that of waiting, delaying the utterance accordingly. While the speech output is in progress, the influence diagram is again monitored so that the speech can be interrupted if waiting becomes more lucrative.

If an agent gets interrupted before speaking the MNI, it will try to repeat the sentence at the next opportunity, unless the interrupting contribution contained the same MNI (compare section 3.3). This way our system can handle disruptive interruptions as well as cooperative completions of the agent's utterance by the interlocutor.

Figure 3 shows two example timelines for the following exchange between the agent and the simulated user (MNI marked with *):

1a) *User: "Uh, I don't know*..."*
1b) *User: "Actually I'm quite satisfied with my old vacuum cleaner."*
2a) *Agent: "Believe me."*
2b) *Agent: "Compared to our Slurp 380* your old vacuum cleaner will look like a stoneage relic."*
3a) *User: "No thank you*."*
3b) *User: "I'm not interested."*
4) *Agent: "The Slurp 380* is the world's first vacuum cleaner with the revolutionary Piranhanado technology!"*
5) *User: "I told you I'm not interested!"*

In both examples, the agent starts phrase 2a after hearing the MNI in phrase 1a. Likewise, the user answers with 3a after the MNI in 2b. Since the introverted agent in the bottom timeline waits for some time after the end of 1b, both 2b and 3a are delayed in this case. 2b is interrupted by 3a (notice the shortened block), but this interruption happend after the MNI was spoken, so the agent will continue with phrase 4 after hearing the MNI in 3a.

5.2 Influence Diagram

We implemented the influence diagram using the SMILE library and the editor GeNIe[2]. It is updated with every new observation about the world state, such as the agent's current role (before the potential behavior change) or voice detection events.

Our current system considers only one input modality, the interlocutor's speech activity. It consists of two separate variables: The speech state, which can be *silent* or *speaking*, and the time since the state's last change. The latter is mapped to the following intervals:

- very short: $duration \in [0.0; 1.0]s$ (backchannels and phrase boundaries)
- short: $duration \in]1.0; 3.0]s$ (short phrases or pauses)
- long: $duration \in]3.0; 5.0]s$ (long pause tolerance [3, 5, 21])
- very long: $duration \in]5.0; \infty]s$ (notably longer pauses, or long-winded speech)

This first influence diagram covers only one single interaction goal, which is to *exert control* over the conversation. We chose this goal because its relationships to Extraversion, Agreeableness and Status are already well documented in literature (see sections 2.2 and 3.4). The utilities were based on those observations about how a character's behavior is perceived. Combinations of situational variables which mark turning points in the behavior were assigned utilities of 1.0 and -1.0, respectively. The values were then extra- and interpolated for the remaining combinations of those variables.

To relate the agent's Status to its personality, the Extraversion and Agreeableness dimensions were uniformly mapped to the range [-1.0; 1.0]. The coordinate space was subdivided further to create 400 samples of different personality configurations. These vectors were then rotated by -37.5° to calculate the corresponding Status values, according to the theory in section 2.1. Finally, the obtained values were mapped to the outcomes of the Status node with their relative frequencies forming the conditional probability distribution.

6 EVALUATION

6.1 Experimental Procedure

We used the described prototype to generate variations of the salesperson conversation from section 5.1. To avoid bias based on the roles in the scenario itself, we removed the semantic content from the sentences. For this the spoken text was scrambled while the MNI bookmarks remained in the same place, in order to create naturally timed but semantically neutral video stimuli. The Extraversion

[2]both by BayesFusion, LLC, and available free of charge for academic teaching and research use at http://www.bayesfusion.com/

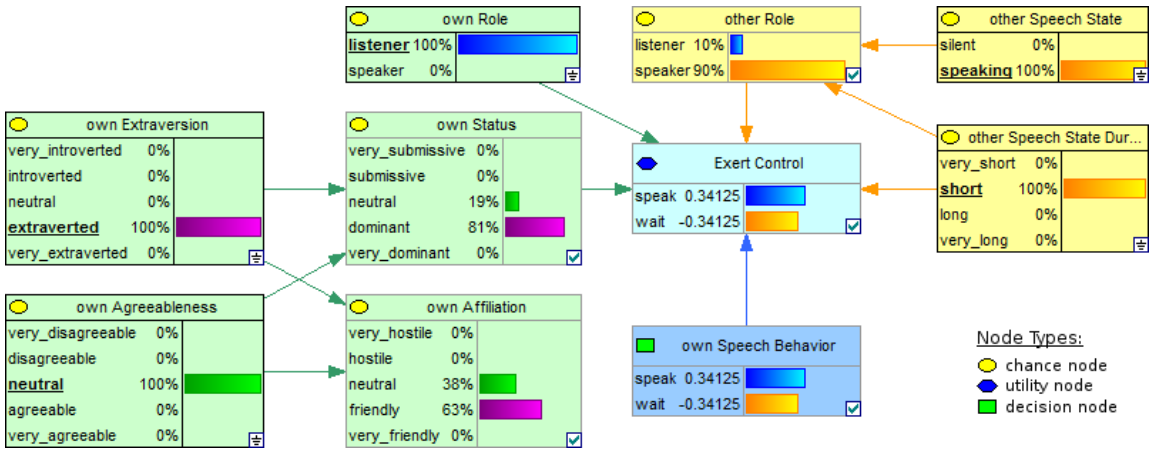

Figure 4: The influence diagram used in our first prototype application. The green nodes on the left describe the agent's state whereas the yellow nodes describe the user's state and observed behavior. The cyan and blue nodes hold the agent's interaction goal and behavior, respectively. The icon in each node's upper left corner shows its type (chance, utility or decision).

of both agents was varied between the levels *introverted* and *extraverted* with Agreeableness kept at *neutral*. This gave us 2×2 video clips with durations between 0:33 and 1:15 minutes.

These video clips were then presented in an online survey where participants had to rate various statements about both agents in each video. The ratings were given on 5-point Likert scales ranging from 1 (disagree completely) to 5 (agree completely). The statements to be rated were taken from the "BFI-S" questionnaire by Lang et al. [15], specifically those for Extraversion and Agreeableness. Two more statements, "has a low rank" and "controls the conversation", were added to measure the Status. For background information, we asked the participants about their age group, gender, first language and prior experience with computer-controlled characters.

Participants were recruited using, among other things, mailing lists, posters and flyers with the survey link. As an incentive for completing the survey, they were invited to enter a lottery afterwards and given the chance to win one of three Amazon gift cards worth 10 Euros.

6.2 Hypotheses

Our hypotheses were as follows:

- Hypothesis 1: An agent's Extraversion score will be higher when it is configured as *extraverted*.
- Hypothesis 2: An agent's Status score will be higher when it is configured as *extraverted*.
- Hypothesis 3: An agent's Agreeableness score will be lower when it is configured as *extraverted*.

Hypothesis 1 was meant to verify that the Extraversion parameter was correctly reflected in the character's speech timing. The other two hypotheses were based on the relationship between the personality and interpersonal stance dimensions (see section 2.1). Therefore, we expected higher Extraversion to imply higher Status as well. Since the timing in our application is only based on the Status and the same Status level can be a result of different personality configurations, we further expected Agreeableness to be affected.

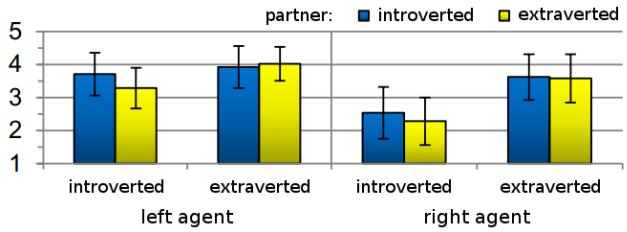

Figure 5: Extraversion scores for the two agents, ranging from 1 (very introverted) to 5 (very extraverted).

6.3 Results

The survey was completed by 116 participants (44 male, 70 female, 2 no answer). The majority (73.3%) was in the age group from 20 to 29, and almost all of them (94%) named German as their first language. Most of the participants (79%) were university students, mainly from subject areas related to computer science or media communications. Therefore, the familiarity with computer-controlled agents was rather high. Most of them had already interacted with video game NPCs and voice assistants, and at least seen social robots in action.

The questionnaire items pertaining to each measured trait - Extraversion, Agreeableness and Status - were combined into a single score for the respective trait, for each agent and condition. 2x2 repeated measures MANOVA were performed to find out whether each agent's configured Extraversion influenced its perceived personality and interpersonal stance. Pairwise comparisons were based on the estimated marginal means with Bonferroni correction.

In the following, *trueEL* will denote the left agent's configured Extraversion while *trueER* will denote that of the right agent.

6.3.1 Perceived Extraversion. We found a significant main effect of *trueEL* on the left agent's perceived Extraversion ($F(1.0, 115.0)=112.97$, $p=0.000$). When set to *extraverted*, it received a higher score ($M=3.97$, $SD=0.58$) than when it was *introverted* ($M=3.50$, $SD=0.67$, $p=0.000$). For the right agent, we found a significant main

Configured Extraversion		Perceived Extraversion			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.71	0.65	2.53	0.79
introverted	extraverted	3.28	0.62	3.62	0.69
extraverted	introverted	3.92	0.64	2.28	0.72
extraverted	extraverted	4.02	0.51	3.58	0.73
Configured Extraversion		**Perceived Status**			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.68	0.69	2.55	0.70
introverted	extraverted	2.61	0.86	3.59	0.83
extraverted	introverted	3.84	0.68	2.46	0.76
extraverted	extraverted	3.44	0.67	2.84	0.83
Configured Extraversion		**Perceived Agreeableness**			
Left Agent	Right Agent	Left Agent		Right Agent	
trueEL	*trueER*	Mean	SD	Mean	SD
introverted	introverted	3.46	0.61	3.48	0.62
introverted	extraverted	3.58	0.69	2.05	0.61
extraverted	introverted	3.39	0.61	3.40	0.59
extraverted	extraverted	2.23	0.54	2.37	0.75

Table 1: Results of the perception study. Perceived traits range from 1.0 (very low) to 5.0 (very high).

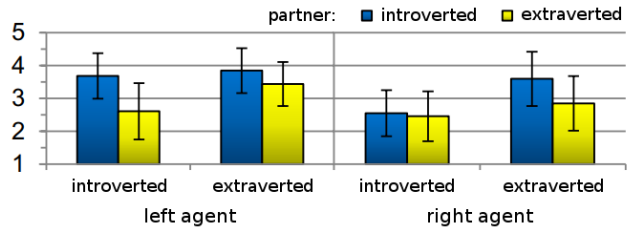

Figure 6: Status scores for the two agents, ranging from 1 (very submissive) to 5 (very dominant).

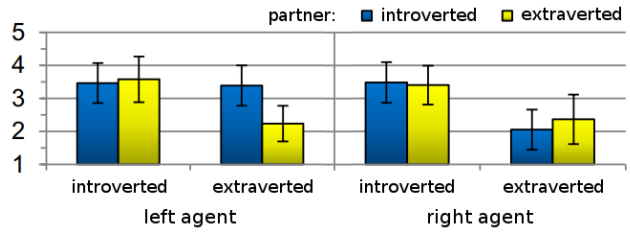

Figure 7: Agreeableness scores for the two agents, ranging from 1 (very disagreeable) to 5 (very agreeable).

effect of *trueER* on its perceived Extraversion (F(1.0, 115.0)=223.13, p=0.000). When set to *extraverted*, it received a higher score (M=3.60, SD=0.06) than when it was *introverted* (M=2.41, SD=0.76, p=0.000). Therefore, Hypothesis 1 was confirmed.

There also was a significant main effect of the left agent's configuration on the score of the right agent (F(1.0, 115.0)=8.17, p=0.005) and vice versa (F(1.0, 115.0)=10.51, p=0.002). In both cases, the difference between the Extraversion scores was more pronounced when the conversational partner was configured as *extraverted*. The effect size was small for the left agent when *trueER* was *introverted* (Cohen's d = ±0.33), but large when *trueER* was *extraverted* (Cohen's d = ±1.29). For the other agent, the effect size was large in both cases (Cohen's d = ±1.46 versus ±1.79)

One possible explanation is that an *extraverted* partner is required in order to see an agent's reaction to being interrupted, which in turn makes it easier to spot the difference in behavior. Overall, the Extraversion score was higher for the left agent, which can be explained by the fact that it always initiated the conversation and, if not interrupted, had more text to say.

6.3.2 Perceived Status. We found a significant main effect of *trueEL* on the left agent's perceived Status (F(1.0, 115.0)=76.01, p=0.000). When set to *extraverted*, it received a higher score (M=3.64, SD=0.71) than when it was *introverted* (M=3.14, SD=0.95, p=0.000). For the right agent, we found a significant main effect of *trueER* on its perceived Status (F(1.0, 115.0)=74.73, p=0.000). When set to *extraverted*, it received a higher score (M=3.22, SD=0.91) than when it was *introverted* (M=2.50, SD=0.73, p=0.000). Therefore, Hypothesis 2 was confirmed.

As with the Extraversion score, we also found a significant main effect of *trueER* on the Status score of the left agent (F(1.0, 115.0)=115.26, p=0.000), and the difference in the score was more pronounced when the other party was *extraverted* (Cohen's d = ±1.08 versus ±0.24). We also found a main effect of *trueEL* on the right agent's Status score (F(1.0, 115.0)=53.01, p=0.000). However, in that case the effect was stronger when the left agent was *introverted* (Cohen's d = ±1.37 versus ±0.49). A reason for this may be that the left agent was perceived as having a higher Status in general, which is in line with its higher Extraversion score and the dependency between those two dimensions (see section 2.1). This in turn could mean that it overshadowed the right agent's behavior differences when it was set to *extraverted*.

6.3.3 Perceived Agreeableness. We found a significant main effect of *trueEL* on the left agent's perceived Agreeableness (F(1.0, 115.0)=182.22, p=0.000). When set to *extraverted*, it received a lower score (M=2.81, SD=0.82) than when it was *introverted* (M=3.52, SD=0.65, p=0.000). For the right agent, we found a significant main effect of *trueER* on its perceived Agreeableness (F(1.0, 115.0)=341.33, p=0.000). When set to *extraverted*, it received a lower score (M=2.21, SD=0.70) than when it was *introverted* (M=3.44, SD=0.60, p=0.000). Therefore, Hypothesis 3 was confirmed.

Again, we found significant main effects of *trueER* on the left agent's Agreeableness score (F(1.0, 115.0)=102.81, p=0.000) and of *trueEL* on the right agent's Agreeableness score (F(1.0, 115.0)=5.36, p=0.022). For the left agent, the effect was only notable when the other party was *extraverted* (Cohen's d = ±2.16 versus ±0.12), whereas for the right agent, the effect was stronger when the left agent was *introverted* (Cohen's d = ±2.34 versus ±1.54). This matches the results for the Status score, and confirms that higher Status implies lower Agreeableness and vice versa (see section 2.1).

6.3.4 Additional Comments. 10 participants made use of the comment fields to state additional observations.

Some described how long the agents paused before speaking and how often they interrupted the other one. They all observed correctly. One person also noticed that the left agent initiated the conversation while the right one merely reacted. As stated above, this may explain the left agent's higher Extraversion and Status.

Others referred to the general tone of voice. One participant found that the agents sounded rather negative, as if they were arguing. In case of the *extraverted* × *extraverted* conversation, another concluded from the monotonous speech that it was not the type of passionate discussion in which overlapping speech was tolerated and even desirable. A different participant wondered whether those two agents were talking to each other or conducting unrelated phone calls, hinting at a lack of interest implied by the overlaps.

Two comments pointed out that aspects such as the degree of control were hard to gauge from the meaningless sounds. One participant gave examples for situations in which overlaps were common and even desirable, which confirms that a believable turn-taking model needs to consider many context factors (see sections 2.2 and 3). We will examine such context factors in future work.

Finally, one person recognized the synthetic voice as that of an American male and drew attention to the strong gender bias induced by this. Said bias was in fact the reason why we had given identical voices to both agents, to reduce the risk that they would be rated according to apparent gender or culture differences. However, we agree that future studies should explore effects of different voices.

7 CONCLUSION

7.1 Summary

We have presented a decision-theoretic approach to the conversational timing and interruption-handling of a computer-controlled agent. Compared to related works, our approach focuses on the personality which the agent is meant to convey, and reasons about explicit goals that depend on the resulting interpersonal stance. For this purpose we gathered relevant findings from psychology and communications research, as well as related approaches to modeling the conversational timing of conversational agents.

Based on those findings, we created an influence diagram in which the agent's desire to control the interaction depends on their personality and subsequently their stance towards the other participant. The latter's intention is inferred from their surface behavior through the Bayesian network. We also described how we integrated such an influence diagram with a dialog system to create a first prototype. In particular, the influence diagram stands between the dialog manager and the agent's behavior realizer, allowing for a distinction between the point in time at which the agent wants to speak, and the one at which its personality dictates that it should.

With this prototype we generated video clips of short agent-agent conversations which we used to validate our model in an online study. The results confirmed that the generated behavior leads to the desired Extraversion perception. Its effects on the perceived Status and Agreeableness are in line with existing psychological theories. However, the survey also confirmed that turn-taking depends on many more factors than those we implemented, which further stresses the need for extensible, adaptable behavior models.

7.2 Open Challenges

There are still many challenges for embedding this model in an interactive application. Most of them fall into the research area of incremental input and output processing.

For instance, the end of the MNI needs to be detected at runtime. For small domains, such as our salesperson dialogue or the "Simon says" game in Chao's work [5], relevant keywords can be marked by hand. Complex domains such as the scenarios in [8, 27] require training on large corpora of user utterances and synonymous phrases. One common approach here is to look for known concepts and entities in the already spoken text and lets the interaction advance once the necessary slots are filled [8, 27, 28].

As for output generation, it is possible that the agent is not yet ready to respond when the turn should be taken. In our prototype, the influence diagram's decision does not force the agent to speak at that point, but rather adds an additional delay if the opposite is true. A suitable extension would be to take the turn and employ turn-hold signals such as filler words [3, 24] and gaze aversion [1, 24] while waiting for the content generation to finish.

Finally, it remains to be seen how well our approach scales with the number of influence factors. Conditional independencies between subsections of the model will help to reduce its complexity, but practical tests are required to assess the performance with increasing numbers of observations, goals and decisions.

7.3 Future Work

Besides speech activity, gaze signals play a vital role in coordinating the conversational behavior and have already been successfully applied to human-robot-interaction [1, 24]. Therefore our next step will be to add this modality to our influence diagram. Additionally, we will explore how the same Bayesian network which recognizes relevant gaze signals can also be used for generation.

So far our network's parameters were calculated from theoretical models or chosen heuristically. To improve its accuracy, we will seek suitable corpora of human-human communication or record such corpora ourselves in order to train the network on realistic data. For instance, we want to model the actual relationship between voice activity and the conversational role to better distinguish between backchannels and barge-in attempts.

We further plan to add more interaction goals, such as "signal involvement", "avoid mistakes" or "save time". For this we will extend our network with chance nodes for semantic information, such as whether the intended speech act is cooperative or disruptive and whether the corresponding MNI was already transmitted (see sections 3.2 and 3.4). Additionally, those goals are likely to depend on other personality factors such as Agreeableness, Neuroticism or Conscientiousness, which we intend to gradually add in order to create a comprehensive behavior model.

ACKNOWLEDGMENTS

The work described in this paper has been partially supported by the BMBF under 16SV7960 within the VIVA project. We would also like to thank everyone who took part in the study or helped with recruiting participants.

REFERENCES

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI '14)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/2559636.2559666

[2] Michael Argyle and Brian R. Little. 1972. Do Personality Traits Apply to Social Behaviour? *Journal for the Theory of Social Behaviour* 2, 1 (1972), 1–33. https://doi.org/10.1111/j.1468-5914.1972.tb00302.x

[3] Dan Bohus and Eric Horvitz. 2011. Decisions About Turns in Multiparty Conversation: From Perception to Action. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI '11)*. ACM, New York, NY, USA, 153–160. https://doi.org/10.1145/2070481.2070507

[4] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 911–920. http://dl.acm.org/citation.cfm?id=2936924.2937059

[5] Crystal Chao. 2015. *Timing multimodal turn-taking in human-robot cooperative activity*. Ph.D. Dissertation. Georgia Institute of Technology.

[6] Cristina Conati. 2013. Virtual Butler: What Can We Learn from Adaptive User Interfaces? In *Your Virtual Butler*, Robert Trappl (Ed.). Lecture Notes in Computer Science, Vol. 7407. Springer-Verlag, Berlin, Heidelberg, 29–41. https://doi.org/10.1007/978-3-642-37346-6_4

[7] Nigel Crook, Cameron Smith, Marc Cavazza, Stephen Pulman, Roger Moore, and Johan Boye. 2010. Handling user interruptions in an embodied conversational agent. In *Proceedings of the AAMAS International Workshop on Interacting with ECAs as Virtual Characters*. Toronto, 27 – 33.

[8] David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse* 2, 1 (2011), 143–170.

[9] Colin G. DeYoung, Yanna J. Weisberg, Lena C. Quilty, and Jordan B. Peterson. 2013. Unifying the Aspects of the Big Five, the Interpersonal Circumplex, and Trait Affiliation. *Journal of Personality* 81, 5 (2013), 465–475. https://doi.org/10.1111/jopy.12020

[10] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 2014. Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 661–668.

[11] Patrick Gebhard, Gregor Mehlmann, and Michael Kipp. 2012. Visual SceneMaker - A Tool for Authoring Interactive Virtual Characters. *Multimodal User Interfaces* 6, 1-2 (2012), 3–11.

[12] Julia A. Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics* 14, 6 (1990), 883 – 903. https://doi.org/10.1016/0378-2166(90)90045-F

[13] Leonard M. Horowitz, Kelly R. Wilson, Bulent Turan, Pavel Zolotsev, Michael J. Constantino, and Lynne Henderson. 2006. How Interpersonal Motives Clarify the Meaning of Interpersonal Behavior: A Revised Circumplex Model. *Personality and Social Psychology Review* 10, 1 (2006), 67–86. https://doi.org/10.1207/s15327957pspr1001_4

[14] Stefan Kopp, Mara Brandt, Hendrik Buschmeier, Katharina Cyra, Farina Freigang, Nicole Krämer, Franz Kummert, Christiane Opfermann, Karola Pitsch, Lars Schillingmann, et al. 2018. Conversational Assistants for Elderly Users–The Importance of Socially Cooperative Dialogue. In *AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications*.

[15] Frieder R. Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G. Wagner. 2011. Short assessment of the Big Five: robust across survey methods except telephone interviewing. *Behavior Research Methods* 43, 2 (June 2011), 548–567. https://doi.org/10.3758/s13428-011-0066-z

[16] RR McCrae and OP John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175.

[17] Robert R McCrae and Paul T Costa. 1989. The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of personality and social psychology* 56, 4 (1989), 586.

[18] Richard E. Neapolitan. 2004. *Learning Bayesian networks*. Pearson Prentice Hall, Upper Saddle River, NJ. OCLC: ocm52534097.

[19] Brian Ravenet, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud. 2015. Conversational Behavior Reflecting Interpersonal Attitudes in Small Group Interactions. In *Intelligent Virtual Agents*, Willem-Paul Brinkman, Joost Broekens, and Dirk Heylen (Eds.). Vol. 9238. Springer International Publishing, Cham, 375–388. https://doi.org/10.1007/978-3-319-21996-7_41

[20] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA.

[21] C. Rich, B. Ponsler, A. Holroyd, and C.L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 375–382. https://doi.org/10.1109/HRI.2010.5453163

[22] William T. Rogers and Stanley S. Jones. 1975. Effects Of Dominance Tendencies On Floor Holding And Interruption Behavior In Dyadic Interaction. *Human Communication Research* 1, 2 (1975), 113–122. https://doi.org/10.1111/j.1468-2958.1975.tb00259.x

[23] Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2013. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics (ACL), 384–393.

[24] Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication* 65 (2014), 50 – 66. https://doi.org/10.1016/j.specom.2014.05.005

[25] Cameron Smith, Nigel Crook, Johan Boye, Daniel Charlton, Simon Dobnik, David Pizzi, Marc Cavazza, Stephen Pulman, Raul Santos de la Camara, and Markku Turunen. 2010. Interaction Strategies for an Affective Conversational Agent. In *Intelligent Virtual Agents*, Jan Allbeck, Norman Badler, Timothy Bickmore, Catherine Pelachaud, and Alla Safonova (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 301–314.

[26] Mark ter Maat, Khiet Phuong Truong, and Dirk K. J. Heylen. 2011. How Agents' Turn-Taking Strategies Influence Impressions and Response Behaviors. *Presence: Teleoperators and Virtual Environments* 20, 5 (Oct. 2011), 412–430. https://doi.org/10.1162/PRES_a_00064

[27] Thomas Visser, David Traum, David DeVault, and Rieks op den Akker. 2012. Toward a model for incremental grounding in spoken dialogue systems. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*.

[28] Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. 2015. An Incremental Turn-Taking Model with Active System Barge-in for Spoken Dialog Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 42–50.